Sentioscope: A Soccer Player Tracking System Using Model Field Particles

Sermetcan Baysal and Pınar Duygulu

***Abstract*— Tracking multiple players is crucial to analyze soccer videos in real time. Yet, rapid illumination changes and occlusions among players who look similar from a distance make tracking in soccer very difficult. Particle-filter-based approaches have been utilized for their ability in tracking under occlusion and rapid motions. Unlike the common practice of choosing particles on targets, we introduce the notion of shared particles densely sampled at fixed positions on the model field. We globally evaluate targets' likelihood of being on the model field particles using our combined appearance and motion model. This allows us to encapsulate the interactions among the targets in the state-space model and track players through challenging occlusions. The proposed tracking algorithm is embedded into a real-life soccer player tracking system called Sentioscope. We describe the complete steps of the system and evaluate our approach on large-scale video data gathered from professional soccer league matches. The experimental results show that the proposed algorithm is more successful, compared with the previous methods, in multiple-object tracking with similar appearances and unpredictable motion patterns such as in team sports.**

***Index Terms*— Model field particles, multiple-object tracking, Sentioscope, soccer player tracking, sports video analysis.**

I. Introduction

SOCCER (football) is among the world's most popular sports played by millions of people around the world. Such popularity has led many computer vision researchers to work on *soccer video analysis*. A wide spectrum of such applications has been introduced to offer team/player performance analysis, referee decision support, video summarization, highlight extraction, and intelligent broadcast cameras [1].

Team/player performance measurement systems has the potential to reveal aspects of the game that are not obvious to the human eye. Such systems can measure the distance covered by players, speed of movement, number of sprints, and players' relative positioning with respect to others. This data are then used in individual player performance evaluation, fatigue detection, assessment of team's tactical performance and analysis of the opponents.

Accurate *tracking of multiple soccer players* in real time is the key issue in performance evaluation, and requires detecting players on video, finding their positions at regular intervals, and linking spatiotemporal data to extract trajectories. However, multiple-player tracking is a nontrivial task due to various challenges. Unlike vehicles or pedestrians, which have relatively predictable motion patterns, soccer players try to confuse each other with unexpected changes in velocity. Moreover, players look almost identical because of their jerseys and they are frequently involved in possession challenges and tackles, where they can be occluded by peer, resulting in tracking ambiguities. Last but not least, environmental conditions can also negatively affect the process of player segmentation. Rapid changes in lighting during cloudy weather cast shadows on the pitch; in sunny weather, dark and long player shadows fall on the field; and in any weather, continuously blinking electronic billboards flash around the stadium. All of these factors can make it difficult to locate players on the field.

It is common to encapsulate the descriptive information of a soccer match (such as player position, velocity and appearance) into states at each time frame to model the game as a collection of temporal states. Then, the multiple-player tracking problem can be perceived as a stochastic process, where the objective is to estimate the state of the game based on the previous observations. Some previous methods use a joint representation of the target space and a unified observation model for all players resulting in a huge state space. A wrong estimation of a single player may negatively affect the whole state and make the formulation intractable. In contrast, other methods decouple the player states and employ separate tracker for each target. Although these approaches are efficient and simpler to formulate, they can neither grasp the global state of the game nor the relations among the players, resulting in the well-known problem of identity hijacking.

As a solution, we propose a robust method to accurately track multiple soccer players in which relative efficiency of employing separate probabilistic trackers is combined with the effectiveness of joint-state models. Players are separately tracked on a model soccer field consisting of shared particles that are densely sampled at unique positions. The overall state of the game and the interactions among the players are encapsulated in the algorithm by globally evaluating the likelihood of the tracks being on the shared particles with respect to our combined appearance and motion model and a color-based occlusion detector. Furthermore, we propose an approach for locating players on the soccer field robust to challenging illumination and environmental conditions.

Manuscript received January 30, 2015; revised May 12, 2015; accepted July 2, 2015. Date of publication July 14, 2015; date of current version July 7, 2016. This work was supported in part by the Sentio Technology and in part by the Bilim Akademisi Genç Bilim İnsanları Ödül Programı. This paper was recommended by Associate Editor H. Wang.

S. Baysal is with the Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: sermetcan@cs.bilkent.edu.tr).

P. Duygulu is with the Department of Computer Engineering, Hacettepe University, Ankara 06800, Turkey (e-mail: pinar@cs.hacettepe.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2015.2455713

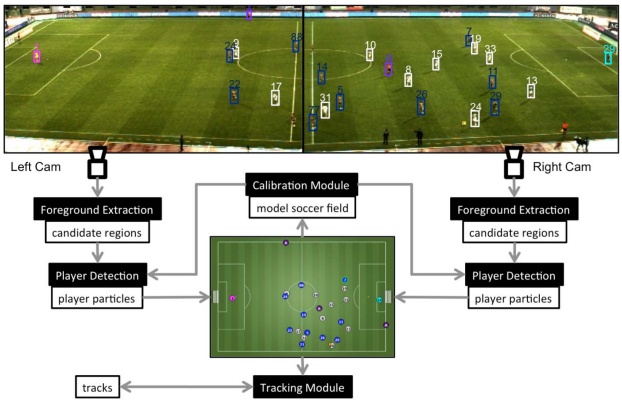

Fig. 1. Overview of Sentioscope: a real-time two-camera soccer player tracking system.

The experimental results demonstrate that our methodology is better at preserving identities of the players during occlusions, and is more suitable for multiple-object tracking with similar appearances such as in team sports compared with the previous methods.

We develop a real-time two-camera soccer-player tracking system called Sentioscope (see Fig. 1), which has successfully tracked players in almost 200 professional soccer league matches and is currently being used in real time as a decision support system that provides data and analytics to the top teams in the Turkish Super League, Italian Serie A, and international competitions.

II. Related Work

The research on target tracking is well rooted and applies to a wide range of domains. Reviewing all studies in the tracking literature is beyond the scope of this paper (see [2] for a detailed survey); thus, in the following, we give a brief review of the studies most relevant to the domain of soccer video analysis.

A. Camera Configuration

One of the most important decisions to make when approaching a sports player tracking problem is camera configuration. In [3]–[9], broadcast footage captured by a pan-tilt-zoom camera is used, offering a relatively cheap and flexible solution to this issue because it is not necessary to physically set up cameras to track players in a game. However, such approaches must deal with continuous changes in view point. A more severe problem is that broadcast videos are usually zoomed to the region of action, and therefore, some players become not visible for tracking. As a solution, studies in [10]–[13] place a number of static cameras to capture a single-view of the entire field. However, as it can be quite challenging for single-view tracking algorithms to resolve frequent and continuous occlusions of players, the methodologies proposed in [14]–[20] tackle the problem by pursuing a multiview approach, in which the observations from four to eight static cameras are fused. Although the efforts of these multiview approaches are laudable, considering the structure of sports arenas/stadiums, these systems introduce extra complications such as difficulties in camera setup, the need to route data to a single processing node, and increased computational complexity, which makes them impractical and relatively expensive for real-time applications.

B. Player Segmentation

Depending on camera configuration, different approaches have been applied for player segmentation. When using static cameras, the simplest way to segment players on the field is to apply background subtraction or statistical background modeling followed by a set of morphological operations, as in [10], [14], [15], and [20]. Background subtraction or modeling is inapplicable if a pan-tilt-zoom camera is being used. Alternatively, assuming color homogeneity of the field, dominant color analysis on a hue channel or histogram back projection can be used to extract a background mask to remove it from the overall image to locate players, as in [3], [5], and [8]. In cases of extreme weather or unstable lighting conditions, these simple player segmentation methods would most likely suffer and generate many false positives. Recently, more sophisticated methodologies have been proposed to cope with such conditions. Gedikli *et al.* [4] employ special templates that extract likelihood maps for player locations based on color distributions, compactness, and vertical spacing cues; Liu *et al.* [6] use a boosted cascade detector using Haar features; Xing *et al.* [9] apply a hybrid multicue learning algorithms with online and offline stages; and Lu *et al.* [7] utilize a deformable part model to automatically locate players.

C. Multiple-Player Tracking

The problem of tracking multiple sports players has been tackled from different perspectives that can be grouped into three main categories.

1) Deterministic Methods: Several approaches employ visual features in a deterministic manner to search for a player's track in the next frame. Color templates are used in early approaches, such as [21]. The idea of kernel density estimation, such as the mean-shift (MS) tracker [22], is applied in [5], using color cues. For better tracking performance, shape information can be decoupled from color, as in [23], or texture and local motion vectors can be used in addition to visual color features, as in [12]. Recent methods such as [24] use a kernelized structured output support vector machine (SVM) to learn the appearance of the track and adapt to changes. To better represent the target and distinguish foreground and background, [25] utilizes a tracking template using discriminative nonorthogonal binary subspace spanned by Haar-like features. Such approaches do not properly encapsulate interactions among players, and therefore, these methods are likely to be distracted when players are occluded or similarly colored tracks are near each other.

2) Data Association and Optimization Based Methods: From another point of view, having detected players in each time unit, one can formulate tracking as a data association problem and seek an optimal solution in a variety of ways.

Gedikli *et al.* [4] use a multiple-hypothesis tracker [26] to create affiliations between current observations and previous player trajectories. A joint probability data association filter [27] is applied to link player observations between consecutive frames in [14] and [20]. Figueroa *et al.* [10] construct a graph in such a way that blobs represent the node edges representing the distance between the blobs, and players are tracked by traversing the graph by considering the minimal path. Shitrit *et al.* [19] formulate a probabilistic occupancy map (POM) of the players as a direct acyclic graph, and find global optimal solution by linear programming. In [28], this time POM is utilized by formulating the problem as a multicommodity network flow. Lu *et al.* [7] use bipartite matching to associate player detections with existing tracks. Liu *et al.* [29] employ hierarchical data association to track sports players with context-conditioned motion models. These approaches require accurate consecutive observations to correctly establish links and theoretically reach a global optimum. Moreover, they involve explicit detection and exhaustive iteration through all associations in a certain time interval, introducing a heavy computational delay that makes them impractical and rather expensive for real-time applications.

3) Probabilistic Methods: The Bayesian framework and its estimations offer another solution to the multiple-player tracking problem. Randomlike movements can be tracked by sequential Monte Carlo estimation, also known as particle filtering [30], which has recently become a popular tracking methodology due to its ability to cope with uncertainties in visual observations and track nonlinear models.

The states of all tracked objects are embodied into a single-joint state and particle filtering techniques are applied for tracking in [31]. This approach was also adopted by Czyz *et al.* [32] for tracking soccer players. The problem with the joint-state model is that it has a size bound; therefore, only a limited number of players can be tracked; more important, inaccuracies in tracking a single player may affect the entire estimation. Several solutions to this problem have been presented, including [6], in which an optimal solution is estimated using a Markov chain Monte Carlo (MCMC) sampler. Collins and Carr [33] proposed a hybrid MCMC algorithm that uses deterministic solutions for blocks of variables to accelerate its stochastic mode-seeking behavior.

Another approach to the player tracking problem is to reduce the state-space size and use separate particle filter trackers for each player, as in [3], [8], [11], [17], [18], and [34]. However, it is crucial for these types of methods to consider players' global state to avoid one player hijacking the track of another due to similar likelihood scores. To cope with this problem, Ok *et al.* [8] use occlusion probability scores; Hess and Fern [34] present discriminative training methods for tracking American football players that attempt to directly optimize the filter parameters in response to observed errors; Kristan *et al.* [11] take advantage of the bird's-eye camera at indoor sports venues and manage the interactions of individual particle filters using a Voronoi partitioning of space.

D. Comparison to Particle Filtering

The particle filtering approaches generate many particles to accurately track each target. Each particle represents a hypothesis for the track, and particles are propagated with respect to an autoregressive model. Pérez *et al.* [35] propose a probabilistic tracker based on particle filters that use similarity of color histograms for likelihood evaluation. To better handle the multimodality of the target distribution that may arise due to the presence of multiple objects, Vermaak *et al.* [36] extend the work of [35] and introduce a mixture particle filter (MPF), in which each object is modeled with an individual particle filter that forms part of a mixture. Okuma *et al.* [37] employ MPF to track hockey players, supported by the Adaboost algorithm [38] for player detection.

The MPF approach performs better than naive particle filtering approaches in resolving basic occlusions among opponents and tracking multiple targets because interactions among the tracks are evaluated by spatially clustering all the particles and allowing particle transfer between different tracks. However, MPF can easily underperform in soccer videos since teammates look almost identical and players are involved in frequent and continuous occlusions. Such cases result in particle degeneration, in which particles of a track are propagated toward another target or transferred to another mixture component. Hence, identity switches or hijackings occur among tracks during full occlusions.

As a solution, instead of employing separate particles for each target, we introduce the idea of densely sampled particles at fixed positions on a model soccer field that are shared among all tracks. Multiple targets are probabilistically tracked on these model field particles, in which the likelihood of a track being on a particle is evaluated globally.

III. Our Approach

A soccer field is modeled using a set of densely sampled particles $\mathbf{S} = \{\mathbf{s}^1, \mathbf{s}^2, \mathbf{s}^3, \ldots, \mathbf{s}^M\}$, where M is the total number of particles needed to span the entire field. These particles discretize the possible position of the players on the model soccer field and each particle $\mathbf{s}^m \in \mathbf{S}$ is represented with a triple, such that $\mathbf{s}^m = \langle \mathbf{q}^m, B^m, A^m \rangle$. The unique 2D position of a sampled particle on the model field is denoted by $\mathbf{q}^m$ and each particle is represented by a corresponding bounding box B^m on the image plane, with an appearance model A^m, as shown in Fig. 2. Bounding boxes overlap with each other on the image plane so that a player always employs a set of neighbor particles. A histogram of oriented gradients (HOG) [39] detector, trained for soccer, is used to decide whether $\mathbf{s}^m \in \mathbf{S}$ contains a player by examining its B^m. It follows that $\mathbf{S^P} \subset \mathbf{S}$ denotes the subset of positively classified particles that are candidates for player positions.

The likelihood of a track/player being on a particle is evaluated by a combination of appearance and motion models. To grasp the global state of the game and the interactions among players, each $\mathbf{s}^m \in \mathbf{S^P}$ is weighted for each track, depending on likelihood scores, and associated with the track having the highest probability. Lower weighted particles are associated with tracks only when a player is completely occluded, which is determined by a simple normal

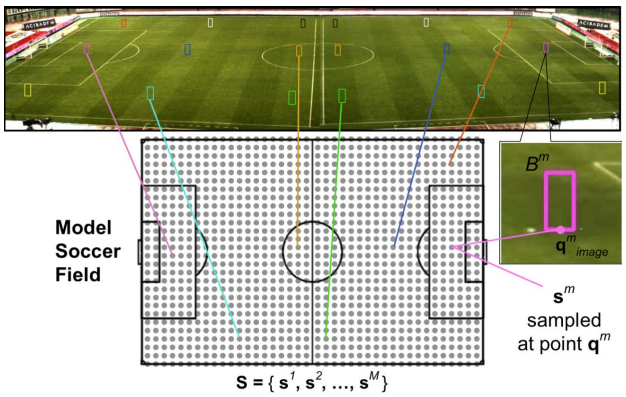

Fig. 2. Soccer field is modeled by densely sampled particles discretizing the possible position of the players. Corresponding bounding boxes for some particles are shown on the image plane. Note that the model field is depicted with sparsely sampled particles for better visualization. In the real case, each square meter contains nine particles.

Bayes classifier [40]. Finally, tracks are separately propagated using a weighted linear combination of their associated particles.

The color and motion models complement each other in multiple-player tracking. Color handles the unpredictable motion patterns since they usually occur when opponents with different colored jerseys are near each other. Motion comes into play when color confuses teammates due to similar appearances. Tactically, teammates show different motion patterns, especially when they are near each other (It is not common for teammates to run side-by-side toward the same direction at same speed). During occlusions, the concept of densely sampled particles and global likelihood (GL) calculation enables players to be aware of each other, keep their locations while their view is blocked.

Fig. 1 shows the global architecture of our system and the interactions among the modules. The following sections describe the details of these components and our work in terms of the model field generation, player segmentation, and multiple-player tracking.

IV. Mapping the Image Plane onto Model Field

We densely sample M particles $\mathbf{S} = \{\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^M\}$ on the model field. Each square meter of the soccer field is spanned by nine sample particles aligned as a 3×3 grid so that a player always stands on many sample particles on the model field. The standard dimensions of the soccer field are 105×68 m, resulting in $M = 64\,260$ particles. In this section, we describe our methods for mapping the captured image onto the model field.

A. Camera Configuration

Our system uses two high-definition Internet protocol cameras to shoot the soccer field; one camera is adjusted to capture the left half and the other is adjusted to capture the right half (see Fig. 1). A narrow portion of the field along the midfield line should be visible in both cameras to establish a homography relation between the tracks in common. The camera synchronization is handled by a software trigger and the exposure is controlled automatically, as in [41], by continuously extracting gray-level histograms of the soccer field, excluding the nonfield regions in the image, and adjusting the exposure until a target mean gray value is obtained.

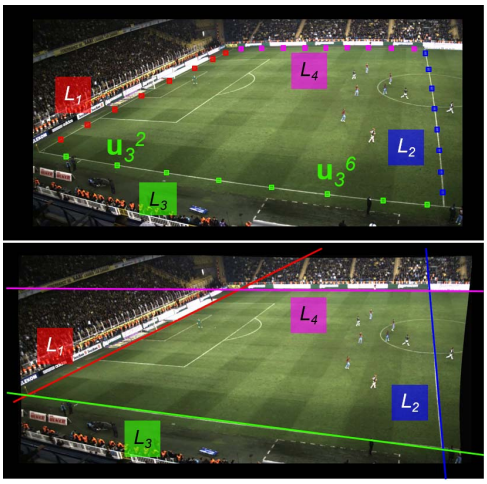

Fig. 3. Top: the calibration points marked on the four field boundary lines on the distorted image. Note the curved appearance of the points on each line. Bottom: the undistorted version of the top image. Observe that the field boundary lines are straight in the undistorted image. L_1: goal line. L_2: midfield line. L_3: near sideline. L_4: far sideline.

B. Distortion Elimination

Since the cameras shoot a large area (68×52.5 m) from close range, the lenses cause radial distortion, resulting in a curved appearance of the actual straight lines in the image. The distortion must be corrected by estimating coefficients, and pixels must be warped to their correct locations. According to Brown's distortion model [42], the relation between a distorted point $\mathbf{q}_d = (x_d, y_d)$ and an undistorted point $\mathbf{q}_{\text{image}} = (x_{\text{image}}, y_{\text{image}})$ on the image plane is expressed as

$$\begin{aligned} x_{\text{image}} &= x_d + (x_d - c_x)\left(K_1 r^2 + K_2 r^4\right) \\ y_{\text{image}} &= y_d + (y_d - c_y)\left(K_1 r^2 + K_2 r^4\right). \end{aligned} \tag{1}$$

Here, K_1 and K_2 are the radial distortion coefficients, (c_x, c_y) are the image center coordinates, and $r = ((x_d - c_x)^2 + (y_d - c_y)^2)^{1/2}$.

The points on the image plane placed on the field boundaries $L_1, \ldots, L_4$ are marked manually (see Fig. 3). These points appear as a curve on the distorted image, but they should form a straight line on the undistorted image. This fact is used to estimate the coefficients by undistorting the marked points using 1 for different values of K_1 and K_2, and choosing the values that minimize the average mean squared error when the lines are fitted to the undistorted points

$$\text{argmin} \sum_{j=1}^{4} \sum_{\mathbf{u}_j^i \in \mathbf{L}_j} \min \|\mathbf{u}_j^i - \mathbf{L}_j\|. \tag{2}$$

Here, $\mathbf{u}_j^i$ is a point marked on the field boundary line j and $\mathbf{L}_j$ is the line fitted to the corresponding set of points.

C. Camera Calibration

Then, the perspective transformation between the particle location points on the model field $\mathbf{q}_{\text{model}}$ and the

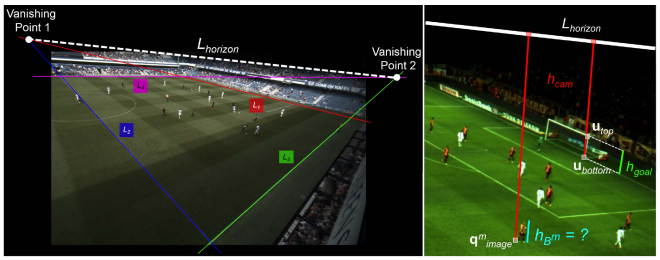

Fig. 4. Calculation of the bounding box height h_{B^m} (in pixels) corresponding to a target height h_T (in meters). A reference object with a known height in meters h_{goal} is utilized to derive the camera height h_{cam}. Then the camera height and the distance to the horizon are used to calculate the height of each bounding box B^m on the image.

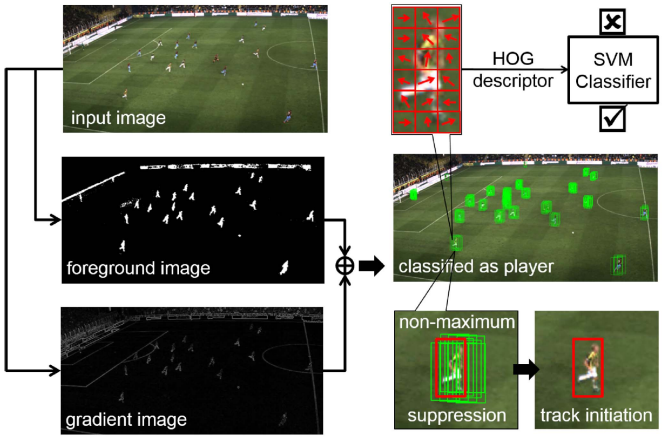

Fig. 5. Steps of player detection and track initiation.

points on the undistorted image plane $\mathbf{q}_{\text{image}}$ are defined as $\mathbf{q}_{\text{image}} = H \cdot \mathbf{q}_{\text{model}}$. (Note that in the following, we refer to $\mathbf{q}_{\text{model}}$ as $\mathbf{q}$ for simplicity.)

Given a set of at least four point correspondences, the homography matrix H can be estimated using direct linear transformation [43]. We use the four corners of the soccer field in the image plane (which are extracted by intersecting the field boundary lines $L_1, \ldots, L_4$) and their correspondences on the model field. Note that more point correspondences can be used to reduce the calibration error.

D. Representing Particles on the Image Plane

Each model field particle $\mathbf{s}^m = \langle \mathbf{q}^m, B^m, A^m \rangle$ is described by its position $\mathbf{q}^m = (x, y)$ and its appearance A^m obtained from the corresponding bounding box B^m, on the image plane. Consider a player over a particle $\mathbf{s}^m$ at position $\mathbf{q}^m$. The corresponding point on the image plane $\mathbf{q}^m_{\text{image}} = H\mathbf{q}^m$ is approximated using perspective transformation, as described in the previous section. Then, the height of B^m, which should be long enough to encapsulate a player, is estimated in pixels to correspond to a fixed height h_T in meters on $\mathbf{q}^m_{\text{image}}$.

The rule of perspectivity states that parallel lines intersect at a vanishing point. As observed in Fig. 4, the line that connects the two vanishing points of the border line pairs (L_1, L_2) and (L_3, L_4) is on the horizon. Since the soccer field is planar, all the imaginary perpendicular lines drawn from the horizon to the soccer field ground in the image plane actually have the same height in the real world, corresponding to the height of the camera above the ground. This principle is utilized to calculate a fixed-height (in meters) bounding box for each model field particle, whereas the bounding box heights in pixels can be different due to the perspective effect. Using the goal posts as reference objects, with a known height of 2.44 m, the bounding box height in pixels for each particle is calculated using a simple direct proportion formula

$$\begin{aligned} h_{\text{cam}} &= (\min\|\mathbf{u}_{\text{bottom}} - L_{\text{horizon}}\| \cdot h_{\text{goal}})/\|\mathbf{u}_{\text{bottom}} - \mathbf{u}_{\text{top}}\| \\ h_{B^m} &= (\min\|\mathbf{q}^m_{\text{image}} - L_{\text{horizon}}\| \cdot h_T)/h_{\text{cam}}. \end{aligned} \tag{3}$$

As vizualized in Fig. 4, here L_{horizon} is the horizon line, $\mathbf{u}_{\text{bottom}}$ and $\mathbf{u}_{\text{top}}$ are the bottom and top of the goal post in the image, h_{goal} is the fixed height of the goal post (equal to 2.44 m) and h_{cam} is the camera height in meters, h_T is a fixed constant for the target bounding box height in meters, and h_{B^m} is the height of the bounding box (in pixels) to be calculated at position $\mathbf{q}^m_{\text{image}}$. The width of B^m is set to half of the height.

V. Player Detection

It is far from reality to expect the standard background subtraction-based approaches to leave only the pixels belonging to the players. In matches played under sunlight or in the absence of sufficient illumination, even shadow detection algorithms are likely to fail at eliminating dark player shadows on the field. Moreover, pixels belonging to the same player may be broken into separate blobs, or a single blob may contain pixels belonging to more than one player. We propose an approach for locating players on the soccer field robust to challenging illumination conditions (see Fig. 5).

A. Foreground Extraction

First, we exploit the foreground segmentation to reduce the number of candidate regions for players. Given an image, the foreground is extracted using the adaptive Gaussian mixture model described in [44] and [45] and is followed by a morphological closing operation for noise removal. Alternative to the fixed global learning rate, we propose using a dynamic spatial learning rate, which is more suitable for soccer videos. The learning rate is automatically adjusted to reconstruct the mixture model if a sudden increase in the number of foreground pixels is detected (indicating a rapid change in lighting). In addition, the learning rates of digital billboard pixels are set to relatively higher values for quick adaptation to continuously changing electronic advertisements.

B. Supervised Player Detection

We aim to decide whether or not a sample particle $\mathbf{s}^m \in \mathbf{S}$ is occupied by a player(s). Since the large number of particles is difficult to exhaustively traverse and process even if it is done in parallel, we reduce the number of model field particles to be examined by extracting the foreground regions, as described. However, during sudden light changes or in presence of dark player shadows, a lot of false positive foreground pixels will

be generated. To ignore the particles with falsely extracted foreground regions, we utilize a classifier for player detection.

We employ the state-of-the-art HOG [39] method for human detection due its abilities to efficiently describe complex shapes and edges in different scales, tolerate small deformations, and cope with illumination and contrast variances.

Recall that each sample particle $\mathbf{s}^m \in \mathbf{S}$ in our model field has a corresponding bounding box B^m as a potential image patch that may encapsulate a player. The bounding boxes with a sufficient ratio of foreground pixels are considered as player candidates and are divided into nonoverlapping 6×3 spatial cells. Gradient orientation histograms, discretizing a range of 0° to 180° into nine bins, are extracted for each spatial region of the candidate B^m. These histograms are then concatenated and normalized to obtain the final HOG descriptor. The HOG descriptors are classified by a linear SVM [46] classifier, trained using a wide spectrum of 60 000 player and 60 000 nonplayer samples collected from over 20 soccer videos with different environmental conditions.

Only the set of positively classified model field particles $\mathbf{S^P} \subset \mathbf{S}$ is used in tracking the players. In the following, for simplicity, we refer to $\mathbf{s}^m$ as a model field particle that is positively classified and discard the particles that are negatively classified. That is, we will only consider $\mathbf{s}^m \in \mathbf{S^P}$. Note that since the operations applied to each sample particle are exactly the same, we distribute the process of player detection to multiple processors.

C. Track Initiation

As observed in Fig. 5, a player stands on many neighbor model field particles with overlapping bounding boxes. To initiate a new track, the overlapping detections are merged using the idea of nonmaximum suppression [47]. A new track is created at a sample particle location with local maximum player detection probability (generated by the SVM classifier). The neighbor particles with overlapping bounding boxes are ignored. Two bounding boxes are said to be overlapping if their geometric centers are closer than some threshold distance. For merging detections along the midfield line, we use plane-to-plane homography to transform geometric centers between images for those bounding boxes that are distributed across different cameras. Note that only those sample particles not occupied by existing players are used in new track initiation.

VI. Multiple-Player Tracking

A. Problem Formulation

A soccer match can be represented by a collection of consecutive states and their forward transitions. The state of the game at any instant can be described using a set of features encapsulating the players' positions, their visual appearances, motion models, and interactions. Then, the objective of tracking multiple players is to estimate the state of the game $\mathbf{x_t}$ at time $\mathbf{t}$, given a set of observations $\mathbf{z_{1:t}}$ up to the present time. If this is assumed to be a first-order Markov process, denoted by $p(\mathbf{x_t}|\mathbf{z_{1:t}})$, then the posterior estimation can be characterized in two steps: 1) involving the prediction of the next state from prior knowledge and 2) performing an update with new observation data [30]

$$p(\mathbf{x_t}|\mathbf{z_{1:t-1}}) = \int p(\mathbf{x_t}|\mathbf{x_{t-1}})\ p(\mathbf{x_{t-1}}|\mathbf{z_{1:t-1}})\ d\mathbf{x_t} \quad (4)$$

$$p(\mathbf{x_t}|\mathbf{z_{1:t}}) \propto p(\mathbf{z_t}|\mathbf{x_t})\ p(\mathbf{x_t}|\mathbf{z_{1:t-1}}). \quad (5)$$

As implied by the prediction (4) and update (5) equations, the posterior estimation process requires specifying the state-space dynamics for describing the state evolution $p(\mathbf{x_t}|\mathbf{x_{t-1}})$ as well as the existence of a model that evaluates the likelihood of an observation for a given state $p(\mathbf{z_t}|\mathbf{x_t})$. We present an efficient and effective estimation of the stochastic process in which each player is represented with a disjoint state and tracked separately. The game's global dynamics and player interactions are captured through the observation model by employing the model field particles as measurements of the states and distributing them among the players using a combined appearance and motion likelihood model.

B. State-Space Dynamics

The state of the game at time $\mathbf{t}$ can be defined as the collection of individual player states $\mathbf{X_t} = \{\mathbf{x_t}^1, \mathbf{x_t}^2, \ldots, \mathbf{x_t}^N\}$, where N is the total number of players/tracks. The state of a player/track $\mathbf{x_t}^n \in \mathbf{X_t}$ is defined as

$$\mathbf{x_t}^n = \begin{bmatrix} \mathbf{p_t}^n & \vec{v}_\mathbf{t}^n & R^n \end{bmatrix} \quad (6)$$

where $\mathbf{p_t}^n = (x, y)$ is the predicted 2D position of the player on the model soccer field, $\vec{v}_\mathbf{t}^n$ is the velocity, and R^n is the reference appearance model of the target being tracked.

1) State Prediction: Omitting the appearance R^n, a Kalman Filter [48] with a constant velocity motion model is used for handling each player state $\mathbf{x_t}^n \in \mathbf{X_t}$, and the prediction of the next state is made as

$$p\left(\mathbf{x_t}^n|\mathbf{x_{t-1}}^n\right) \propto \mathbf{F_t}\ \mathbf{x_{t-1}}^n + \omega_\mathbf{t} \quad (7)$$

where $\mathbf{F_t} = [1\ \ \Delta t;\ \ 0\ \ 1]$ is the state transition model and $\omega_\mathbf{t} \sim N(0, \mathbf{Q})$ is the process noise representing acceleration [which is assumed to be drawn from a zero mean multivariate normal distribution with covariance $\mathbf{Q} = [(\Delta t^4/4)\ \ (\Delta t^3/2);\ \ (\Delta t^3/2)\ \ \Delta t^2]\ \sigma^2_{\text{acc}}$ the acceleration variance and Δt the time between two states expressed in seconds (which is set to 1/frames per second)].

In the following, all the calculations are described for a single time instant $\mathbf{t}$. For simplicity, $\mathbf{x}^n$, $\mathbf{p}^n$, and $\mathbf{s}^m$ denote $\mathbf{x_t}^n$, $\mathbf{p_t}^n$, and $\mathbf{s_t}^m$, respectively.

2) State Update: Recall that the model soccer field $\mathbf{S}$ is spanned by densely sampled particles and a player detector extracts the subset $\mathbf{S^P} \subset \mathbf{S}$ of particles that denote the candidate locations of the tracks on the model field. At each time instant $\mathbf{t}$, $\mathbf{s}^m \in \mathbf{S^P}$ can be represented with the triple $\mathbf{s}^m = \langle \mathbf{q}^m, B^m, A^m \rangle$. Here, $\mathbf{q}^m = (x, y)$ is the fixed 2D location of $\mathbf{s}^m$ on the model soccer field, B^m is the precalculated bounding box in the image plane, and A^m is the current appearance model of the image patch described by B^m.

At each time instant $\mathbf{t}$, the model field particles are distributed among players with respect to the likelihood of track $\mathbf{x}^n$ being on $\mathbf{s}^m$, denoted by $p(\mathbf{s}^m|\mathbf{x}^n)$, which is calculated using a combined appearance and motion model.

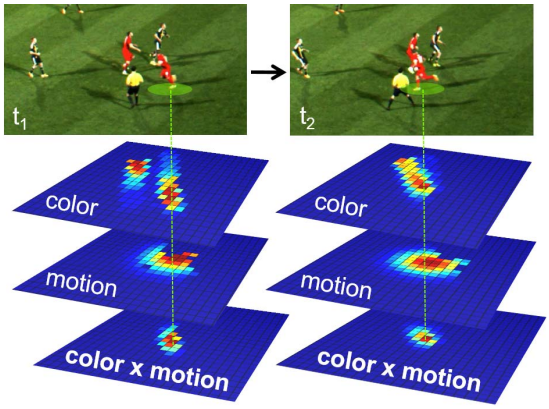

Fig. 6. CML of a tracked player being on the densely sampled model field particles. Likelihood values are normalized and visualized in a jet color map, in which blue and red represents the lowest and highest probabilities, respectively. Observe that the *color x motion* likelihood gives the closest prediction of the player's correct position.

Then the final measurement $\mathbf{p}^n_{\text{observed}} = (x, y)$ of track $\mathbf{x}^n$, indicating the observed position at time $\mathbf{t}$, presumed to be corrupted by a noise ϵ, is calculated as

$$\mathbf{p}^n_{\text{observed}} \sim \sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) \cdot \mathbf{q}^m + \epsilon \tag{8}$$

where $f : \mathbf{X_t} \rightarrow \mathbf{S^P}$ is a functional relation and $f(\mathbf{x}^n) \subset \mathbf{S^P}$ is the subset of sample particles associated with track $\mathbf{x^n}$; $w(\mathbf{x}^n, \mathbf{s}^m)$ is the weight of $\mathbf{s}^m$ extracted by normalizing the likelihood values of track $\mathbf{x}^n$ being on all of the associated particles such that each $w(\mathbf{x}^n, \mathbf{s}^m) \propto p(\mathbf{s}^m|\mathbf{x}^n)$ and $\sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) = 1$; and $\epsilon \sim N(0, \mathbf{Z})$ is the measurement noise, assumed to be a zero mean Gaussian white noise with covariance $\mathbf{Z}$, which is chosen so that the maximum error is about shoulder width (approximately 0.5 m).

Then, given observation $\mathbf{p}^n_{\text{observed}}$ for track $\mathbf{x}^n$, a state update is made using the standard Kalman filter state update equations in [48]. The reference appearance model R^n of the track is extracted from the image patch corresponding to the player's position on the model field and is updated every 10 s to cope with pose and illumination changes.

C. Likelihood Model

The nature of soccer requires the teammates to be spatially separated as much as possible while being as close to their opponents since the opposing teams are involved in possession challenges and tackle with each other. Therefore, color is an important cue to capture the diversity in the appearance of opponents wearing different jerseys. However, utilizing only color features may result in identity hijackings and tracking ambiguities among nearby teammates. As a solution, we propose coupling color features with the target's motion model that yields better tracking of players with similar appearances. The likelihood of a track $\mathbf{x}^n \in \mathbf{X_t}$ being on a particle $\mathbf{s}^m \in \mathbf{S^P}$ at a time $\mathbf{t}$ is separately evaluated for appearance and motion models; then these independent probabilities are multiplied to obtain the overall likelihood, as shown in Fig. 6.

1) Appearance Model: The employed appearance model should be able to handle illumination effects and capture the spatial layout of the color distribution on the players' jerseys. The methods proposed in [35] and [37] are able to successfully cope with such problems. Following these studies, we extract an appearance model for each $\mathbf{s}^m \in \mathbf{S^P}$ by dividing B^m into upper and lower regions and formulating hue–saturation–value (HSV) histograms for each spatial region. An HSV histogram A is composed of a concatenation of separate hue–saturation and value channel histograms, with a total of $C = C_h C_s + C_v$ bins, and $A[c]$ denotes the number of pixels in cth bin, where $c \in \{1, 2, \ldots, C\}$ is the bin index. Each histogram A is normalized to represent the color model as a probability distribution such that $\sum_{c=1}^{C} A[c] = 1$. The reference histogram R^n of each track $\mathbf{x}^n \in \mathbf{X_t}$ is calculated in the same way as the model field particles.

To calculate the color likelihood $p_c(\mathbf{s}^m|\mathbf{x}^n)$, the reference color histogram R^n of track $\mathbf{x}^n$ is compared with the histogram of particle A^m using the Bhattacharyya similarity coefficient. It follows that distance D_{color} between the two color histograms is defined as

$$D_{\text{color}}(R^n, A^m) = \left(1 - \sum_{c=1}^{C} \sqrt{R^n[c] A^m[c]}\right)^{1/2}. \tag{9}$$

It is reported in [35] that successful tracking runs based on color similarity yield consistent exponential behavior for the squared distance D^2_{color}; thus, the color likelihood of a track being on a particle is defined as

$$p_{\text{color}}(\mathbf{s}^m|\mathbf{x}^n) \propto \exp -\lambda \frac{1}{J} \sum_{j=1}^{J} D^2_{\text{color}}(R^n, A^m) \tag{10}$$

where $J = 2$ is the number of subregions (upper and lower body), and R^n and A^m are the color histograms extracted from the subregions of the image patches belonging to $\mathbf{x}^n$ and $\mathbf{s}^m$, respectively. In our experiments, we achieved the best results when the number of bins in the HSV histogram was set to $C_h = 10$ and $C_s = C_v = 5$ when $\lambda = 20$, as in [35].

2) Motion Model: Recall that positional information is maintained by a Kalman Filter and the posterior state of the track is predicted using (7) based on prior knowledge. The motion model evaluates the likelihood of a track $p_{\text{motion}}(\mathbf{s}^m|\mathbf{x}^n)$ by simply measuring the distance D_{motion} between the predicted position of the track and the location of the particle on the model soccer field as

$$D_{\text{motion}}(\mathbf{p}^n, \mathbf{q}^m) = \|\mathbf{p}^n - \mathbf{q}^m\|. \tag{11}$$

Here, $\mathbf{p}^n$ is the predicted position of track $\mathbf{x}^n$ and $\mathbf{q}^m$ is the location of $\mathbf{s}^m$. The motion likelihood is inversely proportional to D_{motion} since it is higher for the particles closer to the predicted position and decreases as the distance between the predicted position and the sampled particle location increases. As a result, the motion likelihood of a track being on a particle, which can be modeled as a normal distribution around the predicted position, is defined using a delta function δ as

$$\delta(d) = \frac{1}{\sigma_{\text{motion}}\sqrt{\pi}} \exp -\frac{d^2}{\sigma^2_{\text{motion}}} \tag{12}$$

$$p_{\text{motion}}(\mathbf{s}^m|\mathbf{x}^n) \propto \delta(D_{\text{motion}}(\mathbf{p}^n, \mathbf{q}^m)). \tag{13}$$

Here, σ_{motion} is the standard deviation of the normal distribution determining the interval of the motion likelihood values. Note that choosing a relatively low σ_{motion} will introduce a larger penalty because the distance between the predicted position and the sampled particle location increases.

3) Combined Appearance and Motion Model: The likelihood of a track $\mathbf{x^n} \in \mathbf{X_t}$ is evaluated separately for the appearance and motion models, using (10) and (13), respectively. Then the overall likelihood is calculated by multiplying the independent probabilities such that

$$p(\mathbf{s^m}|\mathbf{x^n}) \propto p_{\text{color}}(\mathbf{s^m}|\mathbf{x^n}) \cdot p_{\text{motion}}(\mathbf{s^m}|\mathbf{x^n}). \tag{14}$$

Observe in Fig. 6 that motion balances color in the probability multiplication to avoid high likelihood (due to color similarity) between tracks and particles that are far away from each other. This range is controlled by σ_{motion}, which determines the process noise of the motion model and acts as the impact factor of motion on the overall likelihood. A lower value of σ_{motion} will result in dramatically decreasing motion likelihood values as the distance between the predicted position and the particle location increases. In contrast, a higher σ_{motion} narrows down the scale of motion likelihood values and hence increases the impact of color in (14).

D. Global Likelihood Calculation

Although coupling color features with a motion model better represents the target and improves tracking accuracy, there are still many instances in soccer in which the individual player trackers may fail. These occasions include opponents being completely occluded during tackles, teammates standing still near each other so that their similar appearance may result in identity switches, and a bunch of interacting players in challenge of possession during set pieces. To resolve tracking ambiguities in such cases, players' spatial locations with respect to each other should be utilized and the game's global state must be encapsulated in the tracking algorithm. Hence, we propose to distribute the model field particles among the tracks at each instant with respect to the globally calculated likelihoods and estimate the next position of each player using a weighted combination of his associated particles.

At each time instant $\mathbf{t}$, we define a functional relation g:$\mathbf{S}^P \rightarrow \mathbf{X_t}$ where $g(\mathbf{s}^m) \subset \mathbf{X_t}$ denotes the subset of tracks claiming to be on the particle $\mathbf{s}^m$. Each track $\mathbf{x}^n$ claims to be on all nearby particles $\mathbf{s}^m$ such that $\|\mathbf{p}^n - \mathbf{q}^m\| < r_{\max}$, where $r_{\max}$ is the search radius around the predicted position $\mathbf{p}^n$ of track large enough to include the particles that the player can travel in Δt.

When multiple players claim to be on the same particle $\mathbf{s}^m$, the likelihood [calculated using (14)] of each occupying track $\mathbf{x}^n \in g(\mathbf{s}^m)$ is weighted such that the weight $w(\mathbf{s}^m, \mathbf{x}^n) \propto p(\mathbf{s}^m|\mathbf{x}^n)$ and $\sum_{\mathbf{x}^n \in g(\mathbf{s}^m)} w(\mathbf{s}^m, \mathbf{x}^n) = 1$. Then $\mathbf{s}^m$ is only associated with the highest weighed track $\mathbf{x}^* \in g(\mathbf{s}^m)$ and the likelihood value $p(\mathbf{s}^m|\mathbf{x}^*)$ is multiplied with weight $w(\mathbf{s}^m, \mathbf{x}^*)$. While estimating player positions, weight multiplication boosts the probability of those particles having a higher likelihood of a specific track and lower likelihoods for the rest of the tracks. Finally, the observation for a track $\mathbf{x}^n$ is obtained as in (8), using its associated cells $f(\mathbf{x}^n) \subset \mathbf{S^P}$, where $f : \mathbf{X_t} \rightarrow \mathbf{S^P}$ is a functional relation from tracks to sampled particles. The process of the GL calculation and particle distribution is shown in Fig. 7.

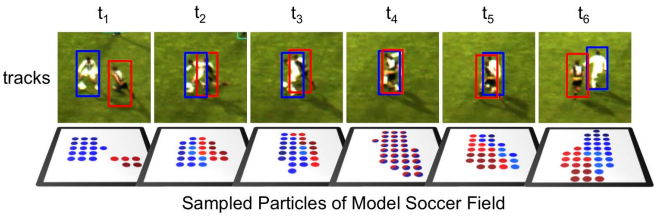

Fig. 7. Top: tracking two players during occlusion. Bottom: the distribution of model field particles between the two tracks with respect to their likelihoods. This figure is best seen in color.

E. Occlusion Handling

Note that a player who is completely occluded by an opponent may have low weights on all nearby sampled particles and hence be lost because none of the particles will be associated with the track. For such situations, we employ a color-based occlusion detector made up of a normal Bayes classifier [40] and associate sampled particles with the tracks under occlusion regardless of their weights. A track $\mathbf{x}^n \in \mathbf{X_t}$ is said to be occluded if its color likelihood $p_{\text{color}}(\mathbf{s}^m|\mathbf{x}^n)$ of all nearby particles are very low values, implying that the view of the track is completely blocked. Therefore, each color likelihood value is evaluated by a normal Bayes classifier, trained on distinct colored jersey samples of different teams, and the track $\mathbf{x}^n$ is classified as occluded or not. The multiplayer tracking methodology described up to this point is set out and summarized in Algorithm 1.

VII. Player Identification

To analyze individual player performances, the identities of the tracks corresponding to real players must be known. Since tracks may be lost and new tracks may be created throughout a game, we employ an optimal assignment-based methodology to automatically maintain identities.

A. Jersey Classification

Based on color, tracks in a soccer match can be classified as belonging to one of five teams or classes: home/away team goalkeeper, home/away team player, and referee. Clusters representing each of these jersey classes are initiated manually by providing a sample color histogram for each class and are updated automatically at regular intervals throughout the game. The reference histogram of a newly created track is compared with those of the samples in each cluster, using the color similarity function in (10) and team identity is assigned using k-nearest-neighbor (k-NN) classification. Note that maintaining clusters with many jersey samples captures the diversity in player appearance due to pose changes and varied illumination in different regions of the field as well as increases classification accuracy.

Algorithm 1 Iteration of Our Multiplayer Tracking Methodology at Time $\mathbf{t}$

Data: Set of model field particles $\mathbf{S^P}$ at $\mathbf{t}$
Result: Update state of each track $\mathbf{x}^n \in \mathbf{X_t}$
foreach $\mathbf{s}^m \in \mathbf{S^P}$ **do** $g(\mathbf{s}^m) \leftarrow \varnothing$;
foreach $\mathbf{x}^n \in \mathbf{X_t}$ **do** $f(\mathbf{x}^n) \leftarrow \varnothing$;
forall $\mathbf{x}^n = \left[\mathbf{p}^n \ \vec{v}^{\ n} \ R^n\right] \in \mathbf{X_t}$ **do**
  $p(\mathbf{x_t}^n|\mathbf{x_{t-1}}^n) \propto \mathbf{F_t}\ \mathbf{x_{t-1}}^n + \omega_\mathbf{t}$;
  forall $\mathbf{s}^m = \langle \mathbf{q}^m, B^m, A^m \rangle \in \mathbf{S^P}$ **do**
    if $\|\mathbf{p}^n - \mathbf{q}^m\| < r_{max}$ **then**
      $p(\mathbf{s}^m|\mathbf{x}^n) \propto p_{color}(\mathbf{s}^m|\mathbf{x}^n) \cdot p_{motion}(\mathbf{s}^m|\mathbf{x}^n)$
      $f(\mathbf{x}^n) \leftarrow \mathbf{s}^m$;
      $g(\mathbf{s}^m) \leftarrow \mathbf{x}^n$;
    end
  end
  if $\forall \mathbf{s}^m \in f(\mathbf{x}^n),\ Bayes\big(p_{color}(\mathbf{s}^m|\mathbf{x}^n)\big) =$ **true** **then**
    $occluded\big(\mathbf{x}^n\big) =$ **true**;
  end
end
forall $\mathbf{s}^m \in \mathbf{S^P}$ **do**
  foreach $\mathbf{x}^n \in g(\mathbf{s}^m)$ **do** $w(\mathbf{s}^m, \mathbf{x}^n) \propto p(\mathbf{s}^m|\mathbf{x}^n)$;
  forall $\mathbf{x}^n \in g(\mathbf{s}^m)$ **do**
    if $w(\mathbf{s}^m, \mathbf{x}^n) = max$ **or** $occluded\big(\mathbf{x}^n\big) =$ **true** **then**
      $p(\mathbf{s}^m|\mathbf{x}^n) \propto p(\mathbf{s}^m|\mathbf{x}^n) \cdot w(\mathbf{s}^m, \mathbf{x}^n)$;
    end
    else
      $f(\mathbf{x}^n) = f(\mathbf{x}^n) - \{\mathbf{s}^m\}$;
    end
  end
end
forall $\mathbf{x}^n \in \mathbf{X_t}$ **do**
  $\mathbf{p}^n_{observed} \sim \sum_{\mathbf{s}^m \in f(\mathbf{x}^n)} w(\mathbf{x}^n, \mathbf{s}^m) \cdot \mathbf{q}^m + \epsilon$
end

B. Assigning Identity Tags to Tracks

Jersey classification is sufficient for maintaining goalkeepers' and the referees' identity tags because they have distinctive jerseys and spatial regions of action on the field. Excluding the goalkeepers, the identity tags of home/away team players are assigned as follows: Let $\mathbf{Y_k} = \{\mathbf{y_k}^1, \mathbf{y_k}^2, \ldots, \mathbf{y_k}^{10}\}$ be the set of real player identities, where $\mathbf{k} \in \{1, 2\}$ indicates the home/away team and let $\mathbf{X_k} \subset \mathbf{X}$ (the time notation is dropped for readability) be the set tracks having the team identity $\mathbf{k}$ assigned by jersey classification. Then, mapping the set of unassigned player tags $\mathbf{Y'_k} \subset \mathbf{Y_k}$ to the set of tracks with no player identities $\mathbf{X'_k} \subset \mathbf{X_k}$ can be formulated as an optimal assignment problem and solved using the Hungarian method [49]. The cost of assigning $\mathbf{y_k}^i \in \mathbf{Y'_k}$ to $\mathbf{x_k}^j \in \mathbf{X'_k}$ is denoted and set as $cost\big(\mathbf{y_k}^i, \mathbf{x_k}^j\big) = \|\mathbf{pp_k}^i - \mathbf{p_k}^j\|$, where $\mathbf{p_k}^j$ is the current track position and $\mathbf{pp_k}^i$ is the estimated position of the unassigned player. The estimated positions are initially set in alignment with the team's tactical lineup and the Hungarian algorithm is run just before kickoff to minimize the overall cost and assign player identity tags to tracks.

TABLE I
ACCURACY OF THE HOG-BASED SVM CLASSIFIER ON PLAYER DETECTION WITH RESPECT TO DIFFERENT CELL CONFIGURATIONS USED DURING FEATURE EXTRACTION

Cell Config.	Feature Size	Accuracy	False Positive
4x2	108	96.69%	2.15%
6x3	360	97.92%	1.24%
8x4	756	97.78%	1.27%

C. Lost Identity Tracking

When a player identity tag $\mathbf{y_k}^i \in \mathbf{Y_k}$ is assigned to a track $\mathbf{x_k}^j \in \mathbf{X_k}$, the estimated position of the player is continuously updated by the track such that $\mathbf{pp_k}^i = \mathbf{p_k}^j$. However, if a player $\mathbf{y_k}^i \in \mathbf{Y'_k}$ is unassigned, then his estimated position $\mathbf{pp_k}^i$ is updated at each time instant by a weighted combination of the positions of the N nearest tracks in $\mathbf{X_k}$ as follows. Let $\{\mathbf{r_k}^1, \mathbf{r_k}^2, \ldots, \mathbf{r_k}^N\}$ be the set of sorted track positions in ascending order with respect to their closeness to $\mathbf{pp_k}^i$, then $\mathbf{pp_k}^i = \sum_{n=1}^{N} w(n)\ \mathbf{r_k}^n$, where $w(n) \propto 1/\|\mathbf{pp_k}^i - \mathbf{r_k}^n\|$ is the weight of the track and $\sum_{n=1}^{N} w(n) = 1$.

The Hungarian algorithm is run throughout the game whenever $|\mathbf{Y'_k}| > 0$ and $|\mathbf{X'_k}| \geq |\mathbf{Y'_k}|$. However, note that when Sentioscope tracks players in a real soccer match and extracting data for professional teams, the process of assigning player identity tags to tracks is supervised by a human operator in real time to correct possible system mistakes.

VIII. EXPERIMENTAL EVALUATION

In this section, we present the experimental results to evaluate our approach and compare it with the baseline and state-of-the-art (SoA) tracking methods.

A. Evaluation of Player Detection Methodology

To evaluate the accuracy of the player detection methodology presented in Section V, we gathered a data set consisting of 60 000 player and 60 000 nonplayer image patches from over 20 different soccer match videos. We then applied ten-fold cross-validation in which 90% of the data set was used for training the HOG-based linear SVM player classifier and the remaining samples were used for testing. Local contrast normalization was made by shifting a larger block of 2×2 spatial cells over the detection window. The best results were obtained when SVM was trained with $C = 1$.

Observe in Table I that a high binary classification accuracy (**97.92%**) is achieved when the bounding box B^m of each sampled particle $\mathbf{s}^k \in \mathbf{S}$ is divided so that it contains 6×3 spatial cells. Note that a low false positive rate can also be tolerated by our tracking methodology during run time. False alarms generated at an unrealistic distance away from the players are naturally eliminated by the tracker. Moreover, some of the false alarms that are generated close to the existing tracks would possibly get low color and motion likelihood (CML) scores and hence cannot disrupt the tracks.

B. Evaluation of Multiple-Player Tracking Algorithm

1) Data sets: We evaluated our proposed multiplayer tracking algorithm using 150 s of the Turkish Super League soccer match played between Istanbul rivals Besiktas and Fenerbahce on April 20, 2014. All players were annotated manually in 5 frames/s, resulting in 750 frames of ground-truth tracking data. The video was captured using two Sentioscope cameras (as described in Section IV-A) at the beginning of the second half of the game, including challenging instances of 55 partial occlusions and 10 full occlusions.

We compare our approach with the other tracking methods using the publicly available Institute of Intelligent Systems for Automation (ISSIA) data set [50]. It consists of 3000 frames captured by six cameras at 25 frames/s, placed around a stadium in a multiview configuration. Furthermore, to evaluate the proposed method on the large scale, we experiment on 10 full-length soccer matches in which the cameras were placed at different heights above the ground. As will be discussed in Section VIII-B7, the error metrics are approximated for this semiannotated tracking data having a total duration of 900 min.

2) Evaluation Criteria: We use the three components of global multiple-object tracking accuracy (GMOTA) [28] (the extended version of MOTA [51]) to measure and evaluate our tracker's ability. As discussed in [19] and [28], GMOTA is more suitable for evaluating sports player tracking where identity preserving is crucial. The false negative (FN), false positive (FP), and global identity miss match (gmme) metrics are defined as follows: $\text{FN} = \sum_t (m_t/g_t)$, $\text{FP} = \sum_t (\text{fp}_t/g_t)$, and $\text{gmme} = \sum_t (\text{gmme}_t/g_t)$. Here, g_t is the number of ground-truth detections, m_t is the number of misses, fp_t is the number of false positives, and gmme_t is the number of identity switches in a frame. Note that lower values are better for these metrics.

Initially, player identity tags are manually assigned to the tracks. Then, in each frame, m_t is incremented for each missing player, both m_t and fp_t are incremented for each player if the distance between a player's ground truth and the observation is more than some threshold (1 m in our results), and gmme_t is incremented for each player whose identity label contradicts with the ground-truth identity.

3) Evaluation of Color and Motion Likelihood: As for the baseline of the proposed multiplayer tracking algorithm, which will be referred to as `CML`, we limit the method to only the `CML` function that combines color and motion features as described in Section VI-C3; with GL calculation (see Section VI-D), occlusion handling (OH) (see Section VI-E), and lost identity tracking (IT) (see Section VII-C) steps omitted. We test the performance of the separate likelihood functions for color and motion and find a suitable σ_{motion} for combining the two features. As observed in Fig. 8, using only color results in more identity switches, whereas using only motion causes a lot of track losses. When color and motion features are combined using a σ_{motion} between 1.25 and 1.5, the tracking errors significantly decrease.

4) Evaluation of Global Likelihood, Occlusion Handling and Identity Tracking: To better understand the effect of `GL` (Section VI-D), `OH` (Section VI-E), and `IT` (Section VII-C) steps, we evaluate the performance over the baseline likelihood function `CML`. As observed in Fig. 9, both `GL` and `OH` resolve more number of occlusions so that FN and gmme is lower than `CML`. When `GL` and `OH` steps are utilized together over `CML` (referred as `Sentioscope`), players are successfully tracked preserving their identities in more instances when they are involved in tackles and challenges. The failure scenarios of the proposed tracking methodology include challenging cases of three or more players involved in a possession or a tackle in which more than one player may be completely occluded. In such cases, the motion model may perform poorly, resulting in track losses. To recover from such cases, the positions of the lost players are continuously estimated and assigned to a newly created track, as described in Section VII. When `IT` step is added, observe that all the error metrics are further reduced below 0.01.

5) Comparison With Baseline Tracking Methods: We compare `Sentioscope` with the baseline tracking methods on our fully annotated data set. OpenCV implementations of MS [22] and optic flow (OF) [52], with the addition of basic OH mechanism similar to the one proposed in [35], were used in the experiments. Color-based particle filter (Color-PF) and color-based MPF (Color-MPF) were implemented with our best effort as described in [35] and [37]. For standardization,

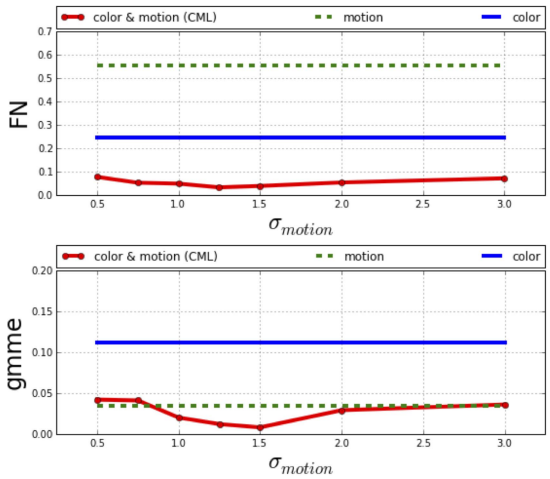

Fig. 8. Evaluation of a likelihood function. The error is significantly lower when color and motion are combined.

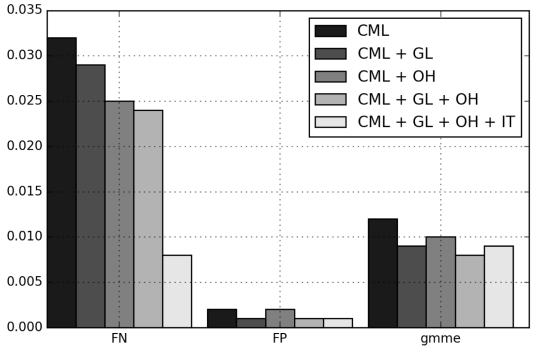

Fig. 9. Changes in FN, FP, and gmme when different steps of our algorithm are added on top of the baseline.

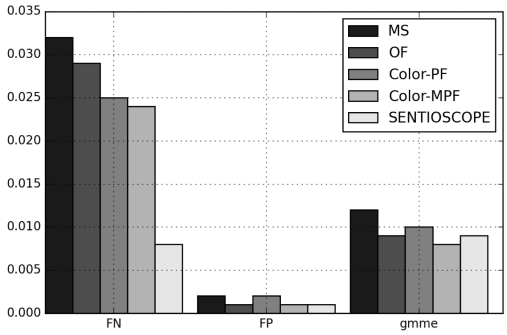

Fig. 10. Sentioscope compared with the baseline tracking methods on our labeled data set.

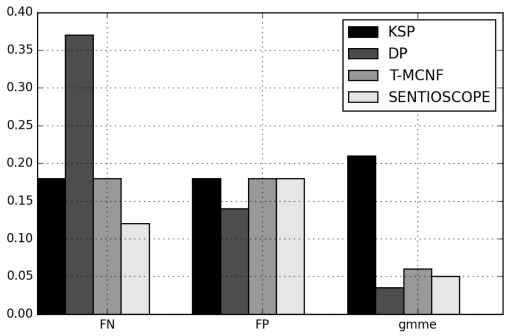

Fig. 11. Sentioscope compared with the multiobject tracking approaches on the ISSIA data set.

player detection and track initiation steps explained in Section V were used in all of the methodologies. Observe in Fig. 10 that particle-filtering-based approaches perform better than the deterministic methods such as OF and MS tracking. Clustering and exchange of particles among the mixtures allows Color-MPF to resolve more occlusions and avoids more identity switches compared with the Color-PF. The results demonstrate that our proposed approach performs better than these baseline methods by achieving a lower FN rate. Since FN is related to the number of missing tracks, the significantly lower FN rate of Sentioscope demonstrates the ability of the method to better track players under occlusion.

6) Comparison With SoA Tracking Methods: We compare Sentioscope with the SoA tracking methods that reported results on the benchmark ISSIA data set [50]. Fig. 11 compares the provided tracking errors in [28] of the K shortest path tracker [53], dynamic program tracker [54], [55], and tracklet-based multicommodity network flow [28] with those of Sentioscope. All of these approaches utilize the multiview configuration of the ISSIA data set. In contrast, we only mapped three cameras (single-view) to our model field and generated field particles as explained in Section IV. Note that there is little overlap between the cameras and some portion of the field is not visible in single-view configuration. This introduces additional errors when tracks are moving between the cameras. Despite the disadvantages, the results show that the concept of model field particles and our proposed recursive tracking algorithm achieves a lower

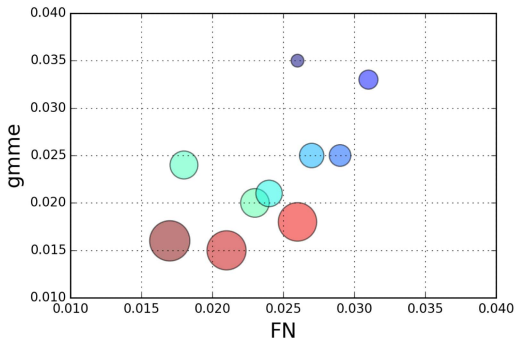

Fig. 12. FN and approximated gmme for 10 different soccer matches. Each colored circle represents a unique 90-minute match. The areas of the circles are proportional to the height of the cameras ranging between 8 and 37 m.

overall error without any time delays as in these methods, which require larger batch sizes for better performance in global trajectory optimization on POMs [54].

7) Evaluation on Semiannotated Data: Frame-by-frame annotation of ground truth tracking data of a soccer video takes days of human effort. However, FN can be exactly computed and an approximation of gmme can be made over full-length games since Sentioscope is embedded in a real-life player tracking system and supervised by a human operator. The job of the operator is to assign player identity tags to new tracks and correct tags throughout the game. The operator interference is utilized to approximate errors from semiannotated data as follows: In each frame, if a player identity tag is not assigned to a track, then m_t is incremented. If the operator corrects the identity tag of a track at time t_n, we assume that the track identity was switched during an occlusion at time t_1. Then, we estimate t_1 and increase gmme_t by $t_n - t_1$.

The scatter plot in Fig. 12 analyzes the changes in FN and gmme with respect to the camera location height. These data are collected from 10 full-length professional soccer matches. Observe that FN (0.017–0.031) and gmme (0.015–0.035) values are similar to those (in Fig. 9) found while experimenting over our data set. This implies that reported results on this data set can be generalized since it reflects the multiple-player tracking challenges encountered in full-length soccer matches. Furthermore, Fig. 12 shows that camera height is influential in the tracking performance of Sentioscope. When cameras are placed to capture the soccer field from 37 m above the ground, occlusions are better handled and low FN and gmme rates are achieved.

C. Evaluation of Jersey/Team Classification

The performance of jersey/team classification is crucial for assigning correct identity tags to the tracks and maintaining them throughout the game, as explained in Section VII. Jersey assignment of a track also reflects the accuracy of the base color likelihood function since the assignment is simply made by comparing the color histogram with the references using (10).

The data set used for the evaluation consists of 15 teams with different jerseys, each with 100 player images in a variety

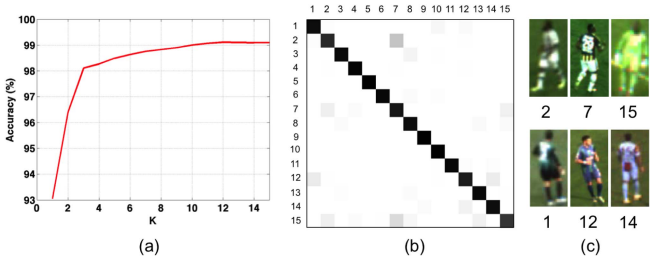

Fig. 13. (a) Accuracy of jersey/team classification with respect to k when k-NN leave-one-out cross-validation is applied. (b) Confusion matrix of jersey classification when $k = 1$ (accuracy is 93%). (c) Top and the bottom rows show the set of most confused team classes in the confusion matrix because of similar jerseys.

TABLE II
RUN TIMES OF THE PLAYER DETECTION AND TRACKING ALGORITHM WITH RESPECT TO THE PARTICLE COUNT

Particle Count	FN	FP	gmme	Run-time (ms)
3x3 per $m^2 \approx 64260$	0.8	0.1	1.3	29.3 ± 25.1
2x2 per $m^2 \approx 28560$	1.2	0.4	1.1	15.8 ± 10.2
1 per $m^2 \approx 7140$	9.7	0.9	7.3	6.6 ± 3.3

of poses. Fig. 13(a) shows the classification accuracy graph for different values of k when k-NN leave-one-out cross-validation is applied. Using $k = 10$ yields a jersey/team classification performance of 99%, which enables us to successfully distinguish the teams in run time and construct the basis for accurately maintaining player identities. Although classification performance is relatively low when $k = 1$, we observe in the confusion matrix in Fig. 13(b) that the majority of the errors are made on jerseys with very similar appearance, but teams would likely not wear these when playing with each other.

D. Computational cost

The algorithm is implemented in C++ with best effort optimization, multithreaded image processing, and GPU usage. The two-camera system runs on a laptop with an Intel i7 CPU with four cores and eight threads at 2.60 GHz. The image acquisition, automatic exposure, light adjustment, foreground extraction, and HOG calculation steps for each camera execute on different threads and have a total run time of 65 ms $\pm$ 6 ms per frame. The player classification and the tracking algorithm, which merge the data from the separate camera threads, execute in parallel. The run times per frame are listed in Table II for different particle configurations.

IX. CONCLUSION

In this paper, we introduce the concept of model field particles that is specifically designed to track interacting players with similar appearances on a precalibrated soccer field plane. Players are tracked through challenging occlusions by utilizing the combined appearance and motion model that is globally evaluated to encapsulate the dynamics of the game. The proposed tracking methodology `Sentioscope` was continuously experimented and evaluated in professional soccer matches throughout its evolution. The experimental results show the effectiveness of our approach compared with other multiple object tracking methods.

REFERENCES

[1] T. D'Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern Recognit.*, vol. 43, no. 8, pp. 2911–2926, 2010.
[2] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, 2006, Art. ID 13.
[3] A. Dearden, Y. Demiris, and O. Grau, "Tracking football player movement from a single moving camera using particle filters," in *Proc. 3rd Eur. Conf. Vis. Media Prod.*, 2006, pp. 29–37.
[4] S. Gedikli, J. Bandouch, N. von Hoyningen-Huene, B. Kirchlechner, and M. Beetz, "An adaptive vision system for tracking soccer players from variable camera settings," in *Proc. 5th ICVS*, 2007.
[5] M.-C. Hu, M.-H. Chang, J.-L. Wu, and L. Chi, "Robust camera calibration and player tracking in broadcast basketball video," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 266–279, Apr. 2011.
[6] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 103–113, 2009.
[7] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1704–1716, Jul. 2013.
[8] H. Ok, Y. Seo, and K. Hong, "Multiple soccer players tracking by condensation with occlusion alarm probability," in *Proc. Workshop Statist. Methods Vis. Process.*, 2002.
[9] J. Xing, H. Ai, L. Liu, and S. Lao, "Multiple player tracking in sports video: A dual-mode two-way Bayesian inference approach with progressive observation modeling," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1652–1667, Jun. 2011.
[10] P. Figueroa, N. Leite, R. M. L. Barros, I. Cohen, and G. Medioni, "Tracking soccer players using the graph representation," in *Proc. 17th ICPR*, 2004, pp. 787–790.
[11] M. Kristan, J. Perš, M. Perše, and S. Kovačič, "Closed-world tracking of multiple interacting targets for indoor-sports applications," *Comput. Vis. Image Understand.*, vol. 113, no. 5, pp. 598–611, 2009.
[12] T. Misu, M. Naemura, W. Zheng, Y. Izumi, and K. Fukui, "Robust tracking of soccer players based on data fusion," in *Proc. 16th ICPR*, 2002, pp. 556–561.
[13] C. J. Needham and R. D. Boyle, "Tracking multiple sports players through occlusion, congestion and scale," in *Proc. BMVC*, vol. 1, no. 1, pp. 93–102, 2001.
[14] S. Iwase and H. Saito, "Parallel tracking of all soccer players by integrating detected positions in multiple view images," in *Proc. 17th ICPR*, 2004, pp. 751–754.
[15] M. Leo, N. Mosca, P. Spagnolo, P. L. Mazzeo, T. D'Orazio, and A. Distante, "Real-time multiview analysis of soccer matches for understanding interactions between ball and players," in *Proc. Conf. Content-Based Image Video Retr.*, 2008, pp. 525–534.
[16] R. Martín and J. M. Martínez, "A semi-supervised system for players detection and tracking in multi-camera soccer videos," *Multimedia Tools Appl.*, vol. 73, no. 3, pp. 1617–1642, 2013.
[17] E. Morais, A. Ferreira, S. A. Cunha, R. M. L. Barros, A. Rocha, and S. Goldenstein, "A multiple camera methodology for automatic localization and tracking of futsal players," *Pattern Recognit. Lett.*, vol. 39, pp. 21–30, Apr. 2014.
[18] E. Morais, S. Goldenstein, A. Ferreira, and A. Rocha, "Automatic tracking of indoor soccer players using videos from multiple cameras," in *Proc. 25th SIBGRAPI*, 2012, pp. 174–181.
[19] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," in *Proc. IEEE ICCV*, Nov. 2011, pp. 137–144.
[20] M. Xu, J. Orwell, and G. Jones, "Tracking football players with multiple cameras," in *Proc. ICIP*, 2004, pp. 2909–2912.
[21] Y. Seo, S. Choi, H. Kim, and K.-S. Hong, "Where are the ball and players? Soccer game analysis with color-based tracking and image mosaick," in *Image Analysis and Processing*. Berlin, Germany: Springer-Verlag, 1997.
[22] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[23] J. Perš and S. Kovačič, "Tracking people in sport: Making use of partially controlled environment," in *Computer Analysis of Images and Patterns*. Berlin, Germany: Springer-Verlag, 2001.
[24] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *Proc. IEEE ICCV*, Nov. 2011, pp. 263–270.
[25] A. Li, F. Tang, Y. Guo, and H. Tao, "Discriminative nonorthogonal binary subspace tracking," in *Proc. 11th ECCV*, 2010, pp. 258–271.

[26] R. L. Streit and T. E. Luginbuhl, "Maximum likelihood method for probabilistic multihypothesis tracking," *Proc. SPIE*, vol. 2235, pp. 394–405, Jul. 1994.
[27] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. San Diego, CA, USA: Academic, 1988.
[28] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Multi-commodity network flow for tracking multiple people," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1614–1627, Aug. 2014.
[29] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 1830–1837.
[30] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
[31] M. Isard and J. MacCormick, "BraMBLe: A Bayesian multiple-blob tracker," in *Proc. 8th IEEE ICVS*, Jul. 2001, pp. 34–41.
[32] J. Czyz, B. Ristic, and B. Macq, "A color-based particle filter for joint detection and tracking of multiple objects," in *Proc. IEEE ICASSP*, Mar. 2005, pp. 217–220.
[33] R. T. Collins and P. Carr, "Hybrid stochastic/deterministic optimization for tracking sports players and pedestrians," in *Proc. 13th ECCV*, 2014, pp. 298–313.
[34] R. Hess and A. Fern, "Discriminatively trained particle filters for complex multi-object tracking," in *Proc. IEEE CVPR*, Jun. 2009, pp. 240–247.
[35] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. 7th ECCV*, 2002, pp. 661–675.
[36] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multimodality through mixture tracking," in *Proc. 9th IEEE ICCV*, Oct. 2003, pp. 1110–1116.
[37] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proc. 8th ECCV*, 2004, pp. 28–39.
[38] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Dec. 2001, pp. I-511–I-518.
[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 886–893.
[40] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA: Academic, 1990.
[41] N. Nourani-Vatani and J. M. Roberts, "Automatic camera exposure control," in *Proc. Austral. Conf. Robot. Autom.*, 2007, pp. 1–6.
[42] D. C. Brown, "Decentering distortion of lenses," *Photogram. Eng.*, vol. 32, no. 3, pp. 444–462, 1966.
[43] A. Agarwal, C. V. Jawahar, and P. J. Narayanan, "A survey of planar homography estimation techniques," Dept. Int. Inst. Inf. Technol., Centre Vis. Inf. Technol., Hyderabad, India, Tech. Rep. 12, 2005.
[44] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. 17th ICPR*, 2004, pp. 28–31.
[45] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, 2006.
[46] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 1997, pp. 130–136.
[47] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th ICPR*, 2006, pp. 850–855.
[48] T. J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 1, pp. 90–99, Jan. 1986.
[49] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.
[50] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P. L. Mazzeo, "A semi-automatic system for ground truth generation of soccer video sequences," in *Proc. 6th IEEE Int. Conf. AVSS*, Sep. 2009, pp. 559–564.
[51] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, May 2008, Art. ID 246309.
[52] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*. Berlin, Germany: Springer-Verlag, 2003.
[53] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.
[54] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, Feb. 2008.
[55] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1201–1208.

Sermetcan Baysal received the B.Sc. and M.Sc. degrees from the Department of Computer Engineering, Bilkent University, Ankara, Turkey, in 2008 and 2011, respectively, where he is currently working toward the Ph.D. degree.

He joined Sentio Technology, Istanbul, Turkey, as the Head of Research and Development to build a commercial product on soccer player tracking and player performance analysis. His academic research is mainly focused on computer vision, particularly, on human action recognition, pedestrian tracking, multiple-object tracking, and sports video analysis.

Pınar Duygulu received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, in 1996, 1998, and 2003, respectively.

She was a Post-Doctoral Researcher with the Informadia Project, Carnegie Mellon University, Pittsburgh, PA, USA. She joined the Department of Computer Engineering, Bilkent University, Ankara, in 2004. From 2014 to 2015, she was with Carnegie Mellon University as a Research Associate. She is currently a Faculty Member with the Department of Computer Engineering, Hacettepe University, Ankara. Her research interests include computer vision and multimedia data mining, in particular, object, face, and action recognition in large image and video collections and analysis of historical documents.